Statistica Neerlandica (2008) Vol. 62, nr. 2, pp. 239–254
doi:10.1111/j.1467-9574.2007.00384.x

Information loss for 2 × 2 tables with missing cell counts: binomial case

Rob Eisinga

Department of Social Science Research Methods, Radboud University Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands

We formulate likelihood-based ecological inference for 2×2 tables with missing cell counts as an incomplete data problem and study Fisher information loss by comparing estimation from complete and incomplete data. In so doing, we consider maximum-likelihood (ML) estimators of probabilities governed by two independent binomial distributions and obtain simplified expressions for their covariance. These expressions reflect well the additional uncertainty arising from the unobserved data compared to complete data tables. We also discuss an approximation to the expected conditional variance of the unobserved entries and ML parameter bias correction. An empirical example is used to demonstrate the results.

Keywords and Phrases: ecological inference, Fisher information loss, expectation–maximization algorithm, missing information principle, parameter bias correction, 2×2 table.

1 Introduction

The ecological inference problem of estimating cell probabilities from the marginal totals of a series of 2×2 contingency tables with missing cell counts has long been attracting the attention of researchers from various disciplines, including political science (King, 1997), sociology (King, Rosen and Tanner, 1999), epidemiology (Salway and Wakefield, 2004), marketing (Böckenholt and Dillon, 2000), econometrics (Golan, Judge and Miller, 1996), agriculture (Magnussen, 2004) and statistics (Plackett, 1977; Hamdan and Nasro, 1986; Haber, 1989; Kocherlakota and Kocherlakota, 1992; McCullagh and Nelder, 1992; Rosen *et al.*, 2001; Wakefield, 2004). While considerable research activity has been directed towards the development of parameter estimation models (see, e.g. King, Rosen and Tanner, 2004), little attention has been given to the key inferential issue of information loss in the maximum-likelihood (ML) estimators relative to those in the complete data setting. Steel, Beh and Chambers (2004) examined the loss of information in 2×2 ecological tables using the observed information matrix for inference and Imai, Lu and Strauss (2007) recently discussed how to quantify information loss using the

r.eisinga@maw.ru.nl

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

missing information principle (MIP) of Orchard and Woodbury (1972). The present paper extends these studies by deriving an analytical expression for the loss of Fisher (expected) information and, correspondingly, the decrement in estimator precision, assuming binomial distributions for the unobserved cell counts. It also discusses an approximation to the expected conditional variance of the missing cell entries to speed up computation for large marginal totals and parameter bias correction for sparse 2×2 ecological tables.

The paper is organized as follows. Section 2 considers the complete and incomplete data likelihoods involved in ecological inference for 2×2 tables and section 3 discusses the application of the expectation–maximization (EM) algorithm and the MIP. Fisher information loss is examined in section 4 and section 5 includes the mean conditional variance approximation and parameter bias correction in ML estimation. An empirical example is offered in section 6 and concluding remarks are given in section 7.

2 Complete and incomplete data likelihoods

We first consider the case in which the interior cells are observed. We assume that the observed data are available in the form of 2×2 contingency tables with completely classified observations from each of S independent units (e.g. groups, areas or other aggregates). The data layout is given in Table 1. The entries, y_{js}, take non-negative integer values and denote the number of observations in row j and column 1, with $j = 0, 1$ and $s = 1, \dots, S$. For the sth table, the row-wise sum of the entries is $[(n_{js} - y_{js}) + y_{js} =] n_{js}$, and the grand total is n_s.

Table 1. Data for table s, $s = 1, \dots, S$.

	$Y=0$	$Y=1$	Total
$X=0$		y_{0s}	n_{0s}
$X=1$		y_{1s}	n_{1s}
Total	$n_s - y_s$	y_s	n_s

The probability distribution of y_{js} obviously depends on the sampling scheme that was used to generate the data. A common model is product binomial sampling, where two independent random samples with fixed n_{js} separately provide estimates of the underlying conditional probabilities $\pi_{0s} = P(Y=1|X=0, s)$ and $\pi_{1s} = P(Y=1| X=1, s)$.

In the complete data setting, the unconditional complete data likelihood for table s is the product of the mutually independent binomial random variables $-B(y_{0s}; n_{0s}, \pi_{0s})$ and $B(y_{1s}; n_{1s}, \pi_{1s})$ – expressed as

$$L_{us}(\pi_{0s}, \pi_{1s}) = P(y_{0s}, y_{1s}|n_{0s}, n_{1s}, \pi_{0s}, \pi_{1s}) = B(y_{0s}; n_{0s}, \pi_{0s})B(y_{1s}; n_{1s}, \pi_{1s})$$
$$= \binom{n_{0s}}{y_{0s}} \pi_{0s}^{y_{0s}}(1-\pi_{0s})^{n_{0s}-y_{0s}} \binom{n_{1s}}{y_{1s}} \pi_{1s}^{y_{1s}}(1-\pi_{1s})^{n_{1s}-y_{1s}}$$

which, as $y_{1s} = y_s - y_{0s}$ and $n_{1s} = n_s - n_{0s}$, equals

$$L_{us}(\pi_{0s}, \pi_{1s}) = \binom{n_{0s}}{y_{0s}} \binom{n_s - n_{0s}}{y_s - y_{0s}} \phi_s'^{y_{0s}} (1 - \pi_{0s})^{n_{0s}} \pi_{1s}^{y_s} (1 - \pi_{1s})^{(n_s - n_{0s}) - y_s} \tag{1}$$

with the parameter $\phi_s' = \pi_{0s}(1 - \pi_{1s})/[\pi_{1s}(1 - \pi_{0s})]$ being the odds ratio for table s. The logarithm of the product binomial likelihood (1) is

$$l_{us}(\pi_{0s}, \pi_{1s}; y_{0s}, y_s) = y_{0s}\phi_s + n_{0s} \log(1 - \pi_{0s}) + y_s \log[\pi_{1s}/(1 - \pi_{1s})] \\ + (n_s - n_{0s}) \log(1 - \pi_{1s}) + \log C_s,$$

where the natural parameter $\phi_s = \log\ \phi_s'$ is the log odds ratio, and

$$C_s = \binom{n_{0s}}{y_{0s}} \binom{n_s - n_{0s}}{y_s - y_{0s}}.$$

This expression shows that the complete data distribution belongs to the exponential family. The log likelihood is linear in ϕ_s and y_{0s} is the only sufficient statistic of the complete data that has to be estimated if the 2×2 cell counts are missing. We also note that for S tables the overall unconditional complete data likelihood is the product of the table-specific product binomial likelihoods, i.e.

$$L_u(\pi_{01}, \ldots, \pi_{0S}, \pi_{11}, \ldots, \pi_{1S}) = \prod_{s=1}^{S} L_{us}(\pi_{0s}, \pi_{1s}) = \prod_s L_{us}(\pi_{0s}) L_{us}(\pi_{1s}),$$

and the overall log likelihood is thus

$$\sum_s l_{us} = \sum_s l_{us}(\pi_{0s}) + \sum_s l_{us}(\pi_{1s}).$$

The ecological inference problem occurs when the row and column totals are observed, but the individual cell entries are unavailable for some reason (e.g. sampling design or data confidentiality). If the row margins n_{js} are fixed, the observed or incomplete data likelihood of the random column margin y_s is the convolution of two binomials, conveniently expressed as an integral of the complete data likelihood

$$L_{os}(\pi_{0s}, \pi_{1s}) = P(y_s | n_{0s}, n_s, \pi_{0s}, \pi_{1s}) = \sum_{i_s = y_{0s}^{l}}^{y_{0s}^{u}} P(i_s, y_s | n_{0s}, n_s, \pi_{0s}, \pi_{1s}),$$

where the integral consists of summing over all possible values of the index frequency $y_{0s}^{l} \le i_s \le y_{0s}^{u}$ given the row and column margins, with lower bound $y_{0s}^{l} = \max [0, y_s - (n_s - n_{0s})]$ and upper bound $y_{0s}^{u} = \min(n_{0s}, y_s)$ (McCullagh and Nelder, 1992). Obviously, the convolution has a binomial distribution with parameters $B(y_s; n_s, \pi_s)$, if $\pi_{0s} = \pi_{1s}$. Using a summary notation such that

$$C_{i_s} = \binom{n_{0s}}{i_s} \binom{n_s - n_{0s}}{y_s - i_s},$$

the convolution log likelihood associated with the observed data for table s is

$$l_{os}(\pi_{0s}, \pi_{1s}; y_s) = \log\left(\sum_{i_s} C_{i_s} e^{i_s \phi_s}\right) + n_{0s}\log(1-\pi_{0s}) + y_s \log[\pi_{1s}/(1-\pi_{1s})]$$
$$+(n_s - n_{0s})\log(1-\pi_{1s}).$$

The conditional likelihood of the complete data for a single table s, conditioning on the column total y_s being fixed, is the non-central hypergeometric distribution obtained as

$$L_{cs}(\pi_{0s}, \pi_{1s}) = P(y_{0s}|y_s, n_{0s}, n_s, \pi_{0s}, \pi_{1s}) = \frac{P(y_{0s}, y_s|n_{0s}, n_s, \pi_{0s}, \pi_{1s})}{P(y_s|n_{0s}, n_s, \pi_{0s}, \pi_{1s})} = \frac{C_s \, e^{y_{0s}\phi_s}}{\sum_{i_s} C_{i_s} \, e^{i_s \phi_s}},$$

which is a function only of the log odds ratio. The log likelihood of the conditional non-central hypergeometric is

$$l_{cs}(\pi_{0s}, \pi_{1s}; y_{0s}) = y_{0s}\phi_s - \log\left(\sum_{i_s} C_{i_s} \, e^{i_s\phi_s}\right) + \log C_s.$$

Note that the complete and incomplete data log likelihoods are related such that

$$l_{os}(\pi_{0s}, \pi_{1s}; y_s) = l_{us}(\pi_{0s}, \pi_{1s}; y_{0s}, y_s) - l_{cs}(\pi_{0s}, \pi_{1s}; y_{0s}), \quad (2)$$

where $l_{us}(\cdot)$ and $l_{cs}(\cdot)$ are the unconditional and the conditional complete data log likelihood for table s, respectively, and $l_{os}(\cdot)$ is the log likelihood of the observed data we wish to maximize in ϕ_s or, equivalently, in the conventional parameters π_{0s} and π_{1s}. Computation of the ML estimates can be achieved using a variety of techniques such as the Newton–Raphson procedure or the Fisher scoring method. We briefly describe the application of the EM algorithm (Dempster, Laird and Rubin, 1977) as it facilitates the discussion of information loss. An obvious identification problem here is that the observed data likelihood contains $2S$ parameters π_{0s} and π_{1s}, to be estimated using only S observed marginal tables. To circumvent this problem, in the remainder we assume – like many statistical procedures for analyzing $2 \times 2 \times S$ contingency tables – that the tables are uniformly homogeneous with respect to the conditional probabilities, i.e. $\pi_{0s} = \pi_0$ and $\pi_{1s} = \pi_1$, for $1 \le s \le S$.

3 Missing information and EM

3.1 Conditional and unconditional expectation

Partially differentiating the identity (2) with respect to ϕ gives the score function

$$\frac{\partial l_{os}}{\partial \phi} = \frac{\partial l_{us}}{\partial \phi} - \frac{\partial l_{cs}}{\partial \phi}. \quad (3)$$

For the 2×2 ecological table s, we have

$$\frac{\partial l_{us}}{\partial \phi} = y_{0s} - n_{0s}\pi_0, \quad \frac{\partial l_{cs}}{\partial \phi} = y_{0s} - \left(\frac{\sum_{i_s} C_{i_s} i_s \, e^{i_s\phi}}{\sum_{i_s} C_{i_s} \, e^{i_s\phi}}\right)$$

and

$$\frac{\partial l_{os}}{\partial \phi} = \left(\frac{\sum_{i_s} C_{i_s} i_s e^{i_s \phi}}{\sum_{i_s} C_{i_s} e^{i_s \phi}} \right) - n_{0s}\pi_0.$$

The observed data log likelihood $l_{os}(\pi_0, \pi_1; y_s)$ does not depend on y_{0s}. Therefore, taking expectations of both sides of (3) over the conditional distribution $P(y_{0s}|y_s, n_{0s}, n_s, \phi)$ or, equivalently, averaging over all possible values of y_{0s} given y_s at a fixed value of ϕ – hereinafter referred to as $E(\cdot|y_s)$ – yields

$$\frac{\partial l_{os}}{\partial \phi} = E\left(\frac{\partial l_{us}}{\partial \phi} \middle| y_s \right) - E\left(\frac{\partial l_{cs}}{\partial \phi} \middle| y_s \right)$$

which, when applied to the 2×2 ecological table s, gives

$$E\left(\frac{\partial l_{us}}{\partial \phi} \middle| y_s \right) = \left(\frac{\sum_{i_s} C_{i_s} i_s e^{i_s \phi}}{\sum_{i_s} C_{i_s} e^{i_s \phi}} \right) - n_{0s}\pi_0,$$

$$E\left(\frac{\partial l_{cs}}{\partial \phi} \middle| y_s \right) = \left(\frac{\sum_{i_s} C_{i_s} i_s e^{i_s \phi}}{\sum_{i_s} C_{i_s} e^{i_s \phi}} \right) - \left(\frac{\sum_{i_s} C_{i_s} i_s e^{i_s \phi}}{\sum_{i_s} C_{i_s} e^{i_s \phi}} \right) = 0,$$

and thus

$$\frac{\partial l_{os}}{\partial \phi} = E\left(\frac{\partial l_{us}}{\partial \phi} \middle| y_s \right). \tag{4}$$

This equation expresses an important relationship in the MIP and EM framework (Woodbury, 1971; Dempster *et al.*, 1977). It states that the score of the observed data log likelihood may be obtained by taking the conditional expectation of the score of the complete data log likelihood given the observed data. Also, as noted by Dempster *et al.* (1977), the first-order derivative of the incomplete data log likelihood with respect to ϕ can be expressed as

$$\frac{\partial l_{os}}{\partial \phi} = E\left(\frac{\partial l_{us}}{\partial \phi} \middle| y_s \right) = E(y_{0s}|y_s, \phi) - E(y_{0s}|\phi).$$

The expectation for the first term on the right is with respect to the complete data given the observed data and the expectation for the second term is with respect to the complete data. For S tables we have

$$U_o(\phi) = U_u(\phi) - U_c(\phi) = \sum_s \left[E(y_{0s}|y_s, \phi) - E(y_{0s}|\phi) \right],$$

where

$$U_{(\cdot)}(\phi) = \sum_s U_{(\cdot)s}(\phi) = \sum_s \partial l_{(\cdot)s} / \partial(\phi).$$

Hence, the total score of the observed data log likelihood is equal to the difference between the conditional expectation of the complete data sufficient statistic y_{0s} given the observed data and the unconditional expectation.

3.2 Expectation–maximization

To specify the EM algorithm when the complete data are from the exponential family distribution, note that if for a single table s we had observed y_{0s}, the ML estimate of ϕ, $\hat{\phi}$, would satisfy $E(y_{0s}|\hat{\phi})=y_{0s}$. Conversely, given some value of ϕ, we may estimate y_{0s} by replacing it by its expected value given y_s and ϕ. For the exponential family density, this is implemented by estimating the complete data sufficient statistic y_{0s} for ϕ by setting it equal to its complete data conditional expectation given y_s and the current guess for ϕ, $\phi^{(t)}$. For S 2×2 ecological tables, the E–step is defined as follows

$$\text{E-step}: y_{0s}^{(t)}=\frac{1}{s}\sum_s\left(\frac{\sum_{i_s} C_{i_s} i_s \mathrm{e}^{i_s\phi^{(t)}}}{\sum_{i_s} C_{i_s}\mathrm{e}^{i_s\phi^{(t)}}}\right)=E(y_{0s}|y_s,\phi^{(t)}).$$

The M-step then takes the estimated complete data and obtains $\phi^{(t+1)}$ by ordinary ML estimation acting as if the estimated data were the observed data. For distributions in the exponential family, actual maximization of the expected log likelihood can be avoided. Instead, the M-step reduces to a simple closed form expression, which for the current problem is

M-step: take

$$\phi^{(t+1)}=\log\left(\frac{\sum_s E(y_{0s}|y_s,\phi^{(t)})\left[\sum_s(n_s-n_{0s})-\left(\sum_s y_s-\sum_s E(y_{0s}|y_s,\phi^{(t)})\right)\right]}{\left[\sum_s y_s-\sum_s E(y_{0s}|y_s,\phi^{(t)})\right]\left[\sum_s n_{0s}-\sum_s E(y_{0s}|y_s,\phi^{(t)})\right]}\right)$$

as the solution to $y_{0s}^{(t)}=E(y_{0s}|\phi)$. Thus, to obtain the next iterate, the conditional expectation of the sufficient statistic y_{0s} computed in the exponential family E-step is substituted for the expected sufficient statistic that occurs in the closed form expression obtained for the complete data ML estimator of ϕ. The E- and M-steps have a likelihood-climbing property and are repeated iteratively until the algorithm converges to the ML estimate of ϕ, which satisfies $E(y_{0s}|\hat{\phi})=E(y_{0s}|y_s,\hat{\phi})$, as the total score $U_\mathrm{o}(\phi)=0$ at $\phi=\hat{\phi}$. Hence, the ML estimate, based on y_s, is the parameter value under which the conditional expectation of y_{0s} given y_s is the same as the unconditional expectation.

3.3 Missing information principle

On differentiating (2) twice with respect to ϕ and multiplying the resulting Hessians by -1, we obtain

$$\frac{-\partial^2 l_{\mathrm{o}s}}{\partial\phi^2}=\frac{-\partial^2 l_{\mathrm{u}s}}{\partial\phi^2}-\frac{-\partial^2 l_{\mathrm{c}s}}{\partial\phi^2}.$$

As the left-hand side is not a function of y_{0s}, taking expectations of both sides over the conditional distribution $P(y_{0s}|y_s,n_{0s},n_s,\phi)$ yields

$$\frac{-\partial^2 l_{\mathrm{o}s}}{\partial\phi^2}=E\left(\frac{-\partial^2 l_{\mathrm{u}s}}{\partial\phi^2}\middle|\,y_s\right)-E\left(\frac{-\partial^2 l_{\mathrm{c}s}}{\partial\phi^2}\middle|\,y_s\right),\tag{5}$$

which, upon setting ϕ to the ML value, is expressed as

$$i_{os}(\phi) = i_{us}(\phi) - i_{cs}(\phi). \tag{6}$$

The term on the left-hand side is the observed information in the observed data and the first term on the right-hand side is the conditional expected value of the complete data information given the observed data. The second term on the right-hand side can be thought of as the information in the unobserved data, conditional on the observed. Thus, (6) gives the fundamental relationship, coined the MIP by Orchard and Woodbury (1972), that the information in the observed data equals the information in the complete data minus the information in the unobserved data. For the current problem, we have

$$\frac{-\partial^2 l_{us}}{\partial \phi^2} = n_{0s}\pi_0(1-\pi_0),$$

$$\frac{-\partial^2 l_{cs}}{\partial \phi^2} = \frac{\sum_{i_s} C_{i_s} i_s^2 \mathrm{e}^{i_s\phi}}{\sum_{i_s} C_{i_s} \mathrm{e}^{i_s\phi}} - \left(\frac{\sum_{i_s} C_{i_s} i_s \mathrm{e}^{i_s\phi}}{\sum_{i_s} C_{i_s} \mathrm{e}^{i_s\phi}}\right)^2,$$

$$\frac{-\partial^2 l_{os}}{\partial \phi^2} = n_{0s}\pi_0(1-\pi_0) - \left[\frac{\sum_{i_s} C_{i_s} i_s^2 \mathrm{e}^{i_s\phi}}{\sum_{i_s} C_{i_s} \mathrm{e}^{i_s\phi}} - \left(\frac{\sum_{i_s} C_{i_s} i_s \mathrm{e}^{i_s\phi}}{\sum_{i_s} C_{i_s} \mathrm{e}^{i_s\phi}}\right)^2\right].$$

As $\partial^2 l_{us}/\partial\phi^2$ is not a function of y_{0s} and $\partial^2 l_{cs}/\partial\phi^2$ is constant for the conditional expectation, taking expectations over the conditional distribution $P(y_{0s}|y_s, n_{0s}, n_s, \phi)$ and setting ϕ to the ML value we have that

$$\begin{aligned} i_{os}(\phi) = \frac{-\partial^2 l_{os}}{\partial \phi^2} &= E\left(\frac{-\partial^2 l_{us}}{\partial \phi^2}\middle| y_s\right) - \left[E\left(\frac{\partial l_{us}}{\partial \phi}^2\middle| y_s\right) - E\left(\frac{\partial l_{us}}{\partial \phi}\middle| y_s\right)^2\right] \\ &= E\left(\frac{-\partial^2 l_{us}}{\partial \phi^2}\middle| y_s\right) - \mathrm{var}\left(\frac{\partial l_{us}}{\partial \phi}\middle| y_s\right). \end{aligned} \tag{7}$$

This result states, as pointed out by Louis (1982), that the observed data observed information can be obtained using the expectations of the derivatives of the complete data likelihood. Also, Dempster *et al.* (1977) have shown that the second-order derivative of the incomplete data log likelihood with respect to ϕ can be expressed as

$$i_{os}(\phi) = \frac{-\partial^2 l_{os}}{\partial \phi^2} = \mathrm{var}(y_{0s}|\phi) - E[\mathrm{var}(y_{0s}|y_s, \phi)].$$

Hence, the observed data observed information can be written as the difference at $\phi = \hat{\phi}$ between the unconditional variance and the conditional variance of the complete data sufficient statistic y_{0s}.

To obtain the expected information, we average both sides of (5) over the marginal distribution $P(y_s|n_{0s}, n_s, \phi)$ of the observed data or, equivalently, over all possible values that $0 \le y_s \le n_s$ can take at a fixed value of ϕ. We denote this expectation operator as $E_{y_s}(\cdot)$. This yields the Fisher information

$$E_{y_s}\left[\frac{-\partial^2 l_{os}}{\partial\phi^2}\right] = E_{y_s}\left[E\left(\frac{-\partial^2 l_{us}}{\partial\phi^2}\middle| y_s\right)\right] - E_{y_s}\left[\mathrm{var}\left(\frac{\partial l_{us}}{\partial\phi}\middle| y_s\right)\right],$$

generally expressed as

$$I_{os}(\phi) = I_{us}(\phi) - E_{y_s}[I_{us}(\phi)]. \tag{8}$$

The asymptotic variance–covariance matrix of the ML parameter estimates is obtained from the inverse of the observed data total expected information $I_o(\hat{\phi}) = \sum_s I_{os}(\hat{\phi})$.

4 Information loss for π_0 and π_1

This section compares the information about the conventional parameters π_0 and π_1 provided by the complete and observed data. If we denote $\boldsymbol{\theta} = (\pi_0\, \pi_1)^{\mathrm{T}}$, the total score vector for S complete data 2×2 tables is

$$\mathbf{U}_u(\boldsymbol{\theta}) = \left[\sum_s \frac{\partial l_{us}}{\partial l\pi_0} \sum_s \frac{\partial l_{us}}{\partial l\pi_1}\right]^{\mathrm{T}} = \begin{bmatrix} \dfrac{y_{0+}}{\pi_0} - \dfrac{n_{0+} - y_{0+}}{1 - \pi_0} \\ \dfrac{y_+ - y_{0+}}{\pi_1} - \dfrac{n_{1+} - (y_+ - y_{0+})}{1 - \pi_1} \end{bmatrix},$$

where the + sign indicates summation over s. The complete data observed information matrix has diagonal elements

$$\mathbf{i}_u(\boldsymbol{\theta})_{\pi_0} = \sum_s \frac{-\partial^2 l_{us}}{\partial \pi_0^2} = \left[\frac{y_{0+}}{\pi_0^2} + \frac{n_{0+} - y_{0+}}{(1-\pi_0)^2}\right],$$

$$\mathbf{i}_u(\boldsymbol{\theta})_{\pi_1} = \sum_s \frac{-\partial^2 l_{us}}{\partial \pi_1^2} = \left[\frac{y_+ - y_{0+}}{\pi_1^2} + \frac{n_{1+} - (y_+ - y_{0+})}{(1-\pi_1)^2}\right]$$

and zero off-diagonal elements as no terms in l_{us} involves both π_0 and π_1, i.e. $\partial^2 l_{us}/\partial\pi_0\,\partial\pi_1 = 0$. As $E(y_{0s}) = n_{0s}\pi_0$, the expected information is

$$\mathbf{I}_u(\boldsymbol{\theta}) = \sum_s E_{y_s}\left(\frac{-\partial^2 l_{us}}{\partial\boldsymbol{\theta}^2}\right) = \mathrm{diag}\left[\frac{n_{0+}}{\pi_0(1-\pi_0)}\ \frac{n_{1+}}{\pi_1(1-\pi_1)}\right],$$

where the expectation is taken over the distribution $P(y_s|n_{0s}, n_s, \pi_0, \pi_1)$. The large sample variance of $\hat{\boldsymbol{\theta}} = (\hat{\pi}_0 \hat{\pi}_1)^{\mathrm{T}}$ is

$$\mathbf{var}_u(\hat{\boldsymbol{\theta}}) = \mathbf{I}_u(\boldsymbol{\theta})^{-1} = \mathrm{diag}\left[\frac{\pi_0(1-\pi_0)}{n_{0+}}\ \frac{\pi_1(1-\pi_1)}{n_{1+}}\right],$$

which is consistently estimated as

$$\widehat{\mathbf{var}}_u(\hat{\boldsymbol{\theta}}) = \mathbf{I}_u(\hat{\boldsymbol{\theta}})^{-1} = \mathrm{diag}\left[\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_{0+}}\ \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1+}}\right].$$

For the observed data, the score vector obtained using (4) is

$$\mathbf{U}_o(\boldsymbol{\theta}) = \left[\sum_s \frac{\partial l_{os}}{\partial l\pi_0} \sum_s \frac{\partial l_{os}}{\partial l\pi_1}\right]^{\mathrm{T}}$$
$$= \begin{bmatrix} \dfrac{\sum_s E(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_0} - \dfrac{n_{0+} - \sum_s E(y_{0s}|y_s,\boldsymbol{\theta})}{1-\pi_0} \\ \dfrac{y_+ - \sum_s E(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_1} - \dfrac{n_{1+} - (y_+ - \sum_s E(y_{0s}|y_s,\boldsymbol{\theta}))}{1-\pi_1} \end{bmatrix},$$

where the expectation is taken over the conditional distribution $P(y_{0s}|y_s, n_{0s}, n_s, \pi_0, \pi_1)$. Using (5) and the identity $\mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta}) = \mathrm{var}(y_{1s}|y_s,\boldsymbol{\theta}) = -\mathrm{cov}[(y_{0s}, y_{1s})|y_s,\boldsymbol{\theta}]$, the observed data observed information matrix has elements

$$i_o(\boldsymbol{\theta})_{\pi_0} = \sum_s \frac{-\partial^2 l_{os}}{\partial \pi_0^2}$$
$$= \left[\frac{\sum_s E(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_0^2} + \frac{n_{0+} - \sum_s E(y_{0s}|y_s,\boldsymbol{\theta})}{(1-\pi_0)^2} - \frac{\sum_s \mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_0^2(1-\pi_0)^2}\right],$$

$$i_o(\boldsymbol{\theta})_{\pi_1} = \sum_s \frac{-\partial^2 l_{os}}{\partial \pi_1^2} = \left[\frac{y_+ - \sum_s E(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_1^2} + \frac{n_{1+} - \left(y_+ - \sum_s E(y_{0s}|y_s,\boldsymbol{\theta})\right)}{(1-\pi_1)^2} - \frac{\sum_s \mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_1^2(1-\pi_1)^2}\right],$$

$$i_o(\boldsymbol{\theta})_{\pi_0\pi_1} = \sum_s \frac{-\partial^2 l_{os}}{\partial \pi_0 \partial \pi_1} = \left[\frac{\sum_s \mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})}{\pi_0(1-\pi_0)\pi_1(1-\pi_1)}\right].$$

As noted by Steel *et al.* (2004), compared to the observed information provided by the complete data, the occurrence of missing cell entries reduces the diagonal elements of the observed data observed information and introduces a positive (variance) term in the off-diagonal elements.

As the expectation of the conditional mean of y_{0s} over the observed data y_s is the expectation of y_{0s} (i.e. $E_{y_s}[E(y_{0s}|y_s,\boldsymbol{\theta})] = E(y_{0s}) = n_{0s}\pi_0$), the expected information matrix obtained using (8) is

$$\mathbf{I}_o(\boldsymbol{\theta}) = \sum_s E_{y_s}\left(\frac{-\partial^2 l_{os}}{\partial \boldsymbol{\theta}^2}\right)$$
$$= \begin{bmatrix} \left[\dfrac{n_{0+}}{\pi_0(1-\pi_0)} - \dfrac{\sum_s E_{y_s}[\mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})]}{\pi_0^2(1-\pi_0)^2}\right] & -\dfrac{\sum_s E_{y_s}[\mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})]}{\pi_0(1-\pi_0)\pi_1(1-\pi_1)} \\ -\dfrac{\sum_s E_{y_s}[\mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})]}{\pi_0(1-\pi_0)\pi_1(1-\pi_1)} & \left[\dfrac{n_{1+}}{\pi_1(1-\pi_1)} - \dfrac{\sum_s E_{y_s}[\mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})]}{\pi_1^2(1-\pi_1)^2}\right] \end{bmatrix}.$$

To account for the difference in expected information in the complete and observed data settings, the diagonal elements of the expected information matrix may be expressed in terms of the off-diagonal elements as

$$I_o(\boldsymbol{\theta})_{\pi_j} = \frac{n_{j+}}{\pi_j(1-\pi_j)}\left[1 - I_o(\boldsymbol{\theta})_{\pi_0\pi_1}\frac{\pi_{1-j}(1-\pi_{1-j})}{n_{j+}}\right],$$

or as

$$I_o(\boldsymbol{\theta})_{\pi_j} = \frac{n_{j+}}{\pi_j(1-\pi_j)}\left[1 - \frac{\sum_s E_{y_s}[\mathrm{var}(y_{js}|y_s,\boldsymbol{\theta})]}{\sum_s \mathrm{var}(y_{js})}\right],$$

where $\mathrm{var}(y_{js}) = n_{js}\pi_j(1-\pi_j)$, $j = 0, 1$. The expectation of the conditional variance of y_{0s} is equal to the unconditional variance of y_{0s} minus the variance of the conditional mean of y_{0s} i.e. $E_{y_s}[\mathrm{var}(y_{0s}|y_s,\boldsymbol{\theta})] = \mathrm{var}(y_{0s}) - \mathrm{var}[E(y_{0s}|y_s,\boldsymbol{\theta})]$. As $n_{j+}[\pi_j(1-\pi_j)]^{-1}$ is the expected information regarding π_j in the complete data, the additional uncertainty introduced by the missing cell counts in the observed data is reflected in the factor

$$R_j(\pi_j, n_{j+}) = 1 - \frac{\sum_s E_{y_s}[\mathrm{var}(y_{js}|y_s,\boldsymbol{\theta})]}{\sum_s \mathrm{var}(y_{js})} = \frac{\sum_s \mathrm{var}_{ys}[E(y_{js}|y_s,\boldsymbol{\theta})]}{\sum_s \mathrm{var}(y_{js})}. \tag{9}$$

Because $\mathrm{var}_{ys}[E(y_{js}|y_s,\boldsymbol{\theta})] \le \mathrm{var}(y_{js})$, the range of $R_j(\pi_j, n_{j+})$ is given by the interval $0 \le R_j(\pi_j, n_{j+}) \le 1$. Hence, the expected information $\mathbf{I}_o(\boldsymbol{\theta})_{\pi_j}$ provided by the observed data approaches the expected information $\mathbf{I}_u(\boldsymbol{\theta})_{\pi_j}$ provided by the complete data, if $\sum_s \mathrm{var}_{ys}[E(y_{js}|y_s,\boldsymbol{\theta})]$ approaches $\sum_s \mathrm{var}(y_{js})$.

The elements of the large sample covariance of $\hat{\boldsymbol{\theta}} = (\hat{\pi}_0\hat{\pi}_1)^{\mathrm{T}}$ in the observed data are

$$\mathrm{var}_o(\hat{\boldsymbol{\theta}})_{\pi_0} = I_o(\boldsymbol{\theta})^{-1}_{\pi_0} = \frac{\pi_0(1-\pi_0)}{n_{0+}}\left[\frac{R_1}{R_0+R_1-1}\right],$$

$$\mathrm{var}_o(\hat{\boldsymbol{\theta}})_{\pi_1} = I_o(\boldsymbol{\theta})^{-1}_{\pi_1} = \frac{\pi_1(1-\pi_1)}{n_{1+}}\left[\frac{R_0}{R_0+R_1-1}\right],$$

$$\mathrm{cov}_o(\hat{\boldsymbol{\theta}})_{\pi_0\pi_1} = I_o(\boldsymbol{\theta})^{-1}_{\pi_0\pi_1} = -\frac{\pi_0(1-\pi_0)}{n_{1+}}\left[\frac{1-R_0}{R_0+R_1-1}\right] = -\frac{\pi_1(1-\pi_1)}{n_{0+}}\left[\frac{1-R_1}{R_0+R_1-1}\right].$$

The two-parameter covariance may alternatively be expressed as

$$\begin{aligned}\mathrm{cov}_o(\hat{\boldsymbol{\theta}})_{\pi_0\pi_1} &= -\left[\frac{\pi_0(1-\pi_0)}{n_{0+}}\left(\frac{1-R_1}{R_0+R_1-1}\right)\frac{\pi_1(1-\pi_1)}{n_{1+}}\left(\frac{1-R_0}{R_0+R_1-1}\right)\right]^{1/2} \\ &= -\left[\Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_0}\Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_1}\right]^{1/2},\end{aligned}$$

where

$$\Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_0} = \mathrm{var}_o(\hat{\boldsymbol{\theta}})_{\pi_0} - \mathrm{var}_u(\hat{\boldsymbol{\theta}})_{\pi_0} \quad \text{and} \quad \Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_1} = \mathrm{var}_o(\hat{\boldsymbol{\theta}})_{\pi_1} - \mathrm{var}_u(\hat{\boldsymbol{\theta}})_{\pi_1}$$

are the variance loss for π_0 and π_1 respectively. Hence, the covariance of π_0 and π_1 is equal to the negative geometric mean of the two variance losses associated with estimating the parameters from the observed data. The correlation between the parameters

$$\mathrm{cor}_o(\hat{\boldsymbol{\theta}})_{\pi_0\pi_1} = -\left[\frac{\Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_0}\Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_1}}{(\mathrm{var}_o(\hat{\boldsymbol{\theta}})_{\pi_0}\mathrm{var}_o(\hat{\boldsymbol{\theta}})_{\pi_1})^{1/2}}\right] = -\left[\frac{(1-R_0)(1-R_1)}{R_0R_1}\right]^{1/2}$$

shows the relation between variance loss and parameter correlation. In the absence of variance loss, the correlation is zero. Conversely, a strong negative correlation is synonymous with large parameter variance increase. A simple measure of parameter collinearity is the variance inflation factor

$$\mathrm{VIF}(\hat{\boldsymbol{\theta}})_{\pi_0\pi_1} = \frac{1}{1-\mathrm{cor}_{\mathrm{o}}(\hat{\boldsymbol{\theta}})^2_{\pi_0\pi_1}} = \frac{R_0R_1}{R_0+R_1-1},$$

with range 1 to infinity.

5 Tables with large and small marginal totals

5.1 Expectation approximation

The expectation of the conditional variance of y_{js} for table s is estimated as

$$E_{y_s}\left[\widehat{\mathrm{var}}(y_{js}|y_s,\hat{\boldsymbol{\theta}})\right] = \sum_{y_s=0}^{n_s}\left(\left[\frac{\sum_{i_s} C_{i_s} i_s^2 \mathrm{e}^{i_s\hat{\phi}}}{\sum_{i_s} C_{i_s}\mathrm{e}^{i_s\hat{\phi}}} - \left(\frac{\sum_{i_s} C_{i_s} i_s \mathrm{e}^{i_s\hat{\phi}}}{\sum_{i_s} C_{i_s}\mathrm{e}^{i_s\hat{\phi}}}\right)^2\right] \times \sum_{i_s} C_{i_s}\mathrm{e}^{i_s\hat{\phi}}(1-\hat{\pi}_0)^{n_{0s}}\hat{\pi}_1^{y_s}(1-\hat{\pi}_1)^{(n_s-n_{0s}-y_s)}\right).$$

The nested summations make this expression cumbersome to compute analytically if n_s is large and the range of possible values for y_{0s} is extensive. There are several ways to speed up computation. One is to use the normal approximation to the binomial convolutional likelihood (Wakefield, 2004) or a saddle point approximation (Davison and Semadeni, 2004). Another option is to approximate the expected value by summing over the distribution of y_s in steps of $w > 1$. If m_s is the integer part of $m_s = n_s w^{-1}$, then

$$E_{y_s}[\widehat{\mathrm{var}}(y_{js}|y_s,\hat{\boldsymbol{\theta}})] \simeq w\sum_{y_s=0}^{m_s}(\cdot),$$

where the constant w is used to re-normalize the probabilities so that they sum to unity. One could also use a conditional sampling procedure to obtain samples from the complete data posterior distribution, filled in via the inverse Bayes formulae (IBF) sampler (Tian, Tan and Ng, 2007). Alternatively, the following approximate relationship in terms of the fitted frequencies may be useful

$$E_{y_s}[\widehat{\mathrm{var}}(y_{js}|y_s,\hat{\boldsymbol{\theta}})] \simeq \left[\frac{1}{n_{0s}\hat{\pi}_0} + \frac{1}{n_{0s}(1-\hat{\pi}_0)} + \frac{1}{n_{1s}\hat{\pi}_1} + \frac{1}{n_{1s}(1-\hat{\pi}_1)}\right]^{-1}. \tag{10}$$

This expression gives a close approximation to the mean conditional variance if $y_{0s}^{\mathrm{u}} - y_{0s}^{\mathrm{l}}$ is large and it is exact if $\hat{\pi}_0 = \hat{\pi}_1$ (i.e. $\hat{\phi}' = 1$). The latter occurs, for example, in studies with a constant ratio, r, of n_{0s} to $(n_s - n_{0s})$ cases, where the total for each table is given by $n_s = n_{0s}(1+r)$. The right-hand side of (10) multiplied by $n_s(n_s-1)^{-1}$

corresponds to McCullagh's (1984) approximation to the conditional cell variance of the non-central hypergeometric distribution.

5.2 Parameter bias correction

Maximum-likelihood estimates are known to be biased when the total Fisher information for the parameters is limited. While for tables with large or moderate-sized counts the amount of bias does not seem to be serious compared to the standard errors, parameter bias can be appreciable in the analysis of a few sparse tables. Let $b(\hat{\pi}_j)$ be the n^{-1} bias of $\hat{\pi}_j$. From the general expression for the biases to order n^{-1} of the ML estimates given by Cox and Snell (1968), we obtain the following expression for the parameter bias in 2×2 ecological tables

$$\begin{aligned} b(\hat{\pi}_j) = \frac{1}{2}[&(\mathbf{I}^{\pi_{jj}})^2(\mathbf{K}_{\pi_j\pi_j\pi_j} + 2\mathbf{J}_{\pi_j\pi_j,\,\pi_j}) + \mathbf{I}^{\pi_{j1-j}}\mathbf{I}^{\pi_{1-j1-j}}(\mathbf{K}_{\pi_{1-j}\pi_{1-j}\pi_{1-j}} + 2\mathbf{J}_{\pi_{1-j}\pi_{1-j},\,\pi_{1-j}}) \\ &+ 2(\mathbf{I}^{\pi_{j1-j}})^2(\mathbf{K}_{\pi_j\pi_{1-j}\pi_{1-j}} + \mathbf{J}_{\pi_j\pi_{1-j},\,\pi_{1-j}} + \mathbf{J}_{\pi_{1-j}\pi_{1-j},\,\pi_j}) \\ &+ \mathbf{I}^{\pi_{jj}}\mathbf{I}^{\pi_{j1-j}}(3\mathbf{K}_{\pi_j\pi_{1-j}\pi_j} + 4\mathbf{J}_{\pi_j\pi_{1-j},\,\pi_j} + 2\mathbf{J}_{\pi_j\pi_j,\,\pi_{1-j}}) \\ &+ \mathbf{I}^{\pi_{jj}}\mathbf{I}^{\pi_{1-j1-j}}(\mathbf{K}_{\pi_j\pi_{1-j}\pi_{1-j}} + 2\mathbf{J}_{\pi_j\pi_{1-j},\,\pi_{1-j}})], \quad j=0,1, \end{aligned} \tag{11}$$

where the superscripts denote matrix inversion of the observed data Fisher information matrix, so that $\mathbf{I}^{\pi_{jr}} = (\mathbf{I}^{-1})_{\pi_{jr}}$, with

$$(\mathbf{I})_{\pi_{jr}} = \sum_s E_{y_s}(-\partial^2 l_{\text{os}}/\partial\pi_j\partial\pi_r), \quad \mathbf{K}_{\pi_r\pi_t\pi_u} = \sum_s E_{y_s}(\partial^3 l_{\text{os}}/\partial\pi_r\partial\pi_t\partial\pi_u) \quad \text{and}$$

$$\mathbf{J}_{\pi_r\pi_t,\,\pi_u} = \sum_s E_{y_s}[(\partial^2 l_{\text{os}}/\partial\pi_r\partial\pi_t)(\partial l_{\text{os}}/\partial\pi_u)],$$

with the expectations taken over the marginal distribution $P(y_s|n_{0s}, n_s, \pi_0, \pi_1)$, and the subscripts r, t and u being replaced by $j=0,1$ when the derivatives are with respect to this parameter.

Under mild regularity conditions the third-order derivatives of the observed data log likelihood satisfy the following identity (McCullagh and Nelder, 1992)

$$\begin{aligned} E_{y_s}\left(\frac{\partial^3 l_{\text{os}}}{\partial\pi_j^3}\right) &= -3\text{cov}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}, \frac{\partial^2 l_{\text{os}}}{\partial\pi_j^2}\right) - E_{y_s}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}\right)^3 \\ &= -3\left[E_{y_s}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}, \frac{\partial^2 l_{\text{os}}}{\partial\pi_j^2}\right) - E_{y_s}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}\right) E_{y_s}\left(\frac{\partial^2 l_{\text{os}}}{\partial\pi_j^2}\right)\right] - E_{y_s}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}\right)^3. \end{aligned}$$

As

$$E_{y_s}(\partial l_{\text{os}}/\partial\pi_j) = E_{y_s}[E(\partial l_{\text{us}}/\partial\pi_j|y_s)] = 0 \quad \text{at} \quad \pi_j = \hat{\pi}_j,$$

the third-order derivatives with respect to π_j are

$$E_{y_s}\left(\frac{\partial^3 l_{\text{os}}}{\partial\pi_j^3}\right) = -3E_{y_s}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}, \frac{\partial^2 l_{\text{os}}}{\partial\pi_j^2}\right) - E_{y_s}\left(\frac{\partial l_{\text{os}}}{\partial\pi_j}\right)^3,$$

$$E_{y_s}\left(\frac{\partial^3 l_{os}}{\partial\pi_j\partial\pi_j\partial\pi_{1-j}}\right) = -2E_{y_s}\left(\frac{\partial l_{os}}{\partial\pi_j}, \frac{\partial^2 l_{os}}{\partial\pi_j\partial\pi_{1-j}}\right) - E_{y_s}\left[\left(\frac{\partial l_{os}}{\partial\pi_{1-j}}\right)\left(\left(\frac{\partial^2 l_{os}}{\partial\pi_j^2}\right) + \left(\frac{\partial l_{os}}{\partial\pi_j}\right)^2\right)\right],$$

where

$$\partial l_{os}/\partial\pi_j = E(\partial l_{us}/\partial\pi_j|y_s),$$
$$\partial^2 l_{os}/\partial\pi_j\partial\pi_{1-j} = E(\partial^2 l_{us}/\partial\pi_j\partial\pi_{1-j}|y_s) - E(\partial l_{us}/\partial\pi_j|y_s)E(\partial l_{us}/\partial\pi_{1-j}|y_s) \quad \text{and}$$
$$\partial^2 l_{os}/\partial\pi_j^2 = E(\partial^2 l_{us}/\partial\pi_j^2|y_s) + E(\partial {l_{us}}^2/\partial\pi_j|y_s) - E(\partial l_{us}/\partial\pi_j|y_s)^2,$$

with the expectations denoted E_{y_s} taken over the marginal distribution $P(y_s|n_{0s}, n_s, \pi_0, \pi_1)$, and the expectations denoted E taken over the conditional distribution $P(y_{0s}| y_s, n_{0s}, n_{1s}, \pi_0, \pi_1)$. Once the biases are determined, the bias-corrected ML estimates $\hat{\pi}_j^c$ can be obtained using $\hat{\pi}_j^c = \hat{\pi}_j - b(\hat{\pi}_j)$, where $\hat{\pi}_j$ are the uncorrected ML estimates. Finally, it can be shown that the parameter biases are related to the parameter variance loss by

$$\Delta\text{var}(\hat{\theta})_{\pi_j} = \text{var}_o(\hat{\theta})_{\pi_j} - \text{var}_u(\hat{\theta})_{\pi_j} = \left[\frac{b(\hat{\pi}_j)}{b(\hat{\pi}_{1-j})}\right]\text{cov}_o(\hat{\theta})_{\pi_j\pi_{1-j}}.$$

The loss in parameter variance in 2×2 ecological tables is equal to the product of the bias ratio and the two-parameter covariance. Hence, the expression states that $|b(\hat{\pi}_j)| > |b(\hat{\pi}_{1-j})|$, if $\Delta\text{var}(\hat{\theta})_{\pi_j} > |\text{cov}_o(\hat{\theta})_{\pi_j\pi_{1-j}}|$.

6 Empirical example

We illustrate our results with a simple aggregate data example assessing the relationship between student-reported course workload and teacher-awarded course grades. The data available for analysis consist of the number of students (n_s), the number of students who reported nominal workload or less in an anonymous course evaluation (n_{0s}) and the number of students who received a mark of eight or higher (y_s), for $S = 6$ courses. The cross-classified counts (e.g. the number of students with both nominal workload or less and mark $\geq$ 8) are unavailable. The total and marginal counts for the $s = (1, \ldots, 6)$ tables are $n = (19, 20, 17, 18, 20, 20)$, $n_0 = (18, 18, 8, 9, 10, 2)$ and $y = (3, 6, 8, 9, 9, 14)$. Table 2 displays the history of the EM iterations.

The ML estimates — given by the EM algorithm as $\hat{\pi} = (0.1395, 0.8014)$ — indicate that the probability of obtaining a mark of eight or higher is much lower among students who reported nominal workload or less, than among those who reported above nominal workload.

The Fisher information matrix provided by the observed data and its inverse, evaluated at the ML estimates, are

$$\mathbf{I}_o(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} 363.8380 & 127.6091 \\ 127.6091 & 217.8595 \end{pmatrix} \quad \text{and} \quad \mathbf{I}_o(\hat{\boldsymbol{\theta}})^{-1} = 10^{-3}\begin{pmatrix} 3.4591 & -2.0261 \\ -2.0261 & 5.7769 \end{pmatrix}.$$

Table 2. Expectation-maximization iterates.

t	$\phi^{(t)}$	$\phi^{(t)}-\hat{\phi}$	$\pi_0^{(t)}$	$\pi_1^{(t)}$	$E(y_0\|y,\phi^{(t)})$	$E(y_0\|\phi^{(t)})$
0	1.000000	4.213950	–	–	4.420392	
1	−0.143280	3.070670	0.414412	0.449553	3.571372	4.420392
2	−0.892954	2.320995	0.334816	0.551435	2.971272	3.571372
3	−1.455839	1.758111	0.278557	0.623447	2.535595	2.971272
4	−1.899475	1.314474	0.237712	0.675729	2.218411	2.535595
5	−2.251035	0.962915	0.207976	0.713791	1.990113	2.218411
10	−3.063948	0.150001	0.148943	0.789353	1.554416	1.588723
25	−3.213695	0.000255	0.139566	0.801356	1.488641	1.488699
50	−3.213950	0.000000	0.139550	0.801376	1.488533	1.488533
52	−3.213950	0.000000	0.139550	0.801376	1.488533	1.488533

The expected information and variance–covariance matrices one would have if the complete data were available are

$$\mathbf{I}_{\mathrm{u}}(\hat{\boldsymbol{\theta}})=\begin{pmatrix}532.9967 & 0\\ 0 & 314.1246\end{pmatrix} \quad\text{and}\quad \mathbf{I}_{\mathrm{u}}(\hat{\boldsymbol{\theta}})^{-1}=10^{-3}\begin{pmatrix}1.8762 & 0\\ 0 & 3.1835\end{pmatrix}.$$

The rate of increase in the asymptotic variance of the parameters due to the missing cell counts can be summarized by the scalar

$$\lambda_j=\mathrm{I}_{\mathrm{o}}(\hat{\boldsymbol{\theta}})^{-1}_{\pi_j}[\mathrm{I}_{\mathrm{u}}(\hat{\boldsymbol{\theta}})^{-1}_{\pi_j}]^{-1}=[R_{1-j}(R_j+R_{1-j}-1)^{-1}],$$

with R_j defined as in (9), $j=0,1$. For $\hat{\pi}_0$ and $\hat{\pi}_1$, we find an increase rate of $\lambda_0=1.8437$ and $\lambda_1=1.8147$ respectively. The parameter covariance is easily verified to equal the geometric mean of the parameter variance losses multiplied by -1, i.e.

$$-[(0.0034591-0.0018762)\times(0.0057769-0.0031835)]^{1/2}=-0.0020261.$$

The parameter correlation is estimated as -0.4533 and the variance inflation factor as 1.2586, implying that in this particular data example collinearity is not severe. The approximate expectation of the conditional variance of the missing cell entries is given in Table 3.

Table 3. Estimated expected conditional variance of y_{js} and its approximation.

s	$E_{y_s}[\widehat{\mathrm{var}}(y_{js}\|y_s,\hat{\boldsymbol{\theta}})]$	$([n_{0s}\hat{\pi}_0(1-\hat{\pi}_0)]^{-1}+[n_{1s}\hat{\pi}_1(1-\hat{\pi}_1)]^{-1})^{-1}$
1	0.1419	0.1482
2	0.3622	0.3870
3	0.5242	0.5750
4	0.5636	0.6160
5	0.6321	0.6844
6	0.2150	0.2216

Although the approximation (10) is not called for in this sparse data example, the figures indicate that even for a small number of 2×2 tables with meagre marginal totals it works reasonably well. The variance loss for $\hat{\pi}_0$ is smaller than the absolute parameter covariance; so, there is less absolute bias in $\hat{\pi}_0$ than there is in $\hat{\pi}_1$. Indeed, the biases of $\hat{\pi}_0$ and $\hat{\pi}_1$ were estimated using (11) as $b(\hat{\pi})=(-0.0051,0.0065)$, and the bias-corrected ML estimates were correspondingly determined as $\hat{\pi}^{\mathrm{c}}=(0.1446,0.7949)$.

The product of the ratio of the biases and the parameter covariance equals the parameter variance loss. For $\hat{\pi}_0$, we have

$$\Delta\mathrm{var}(\hat{\boldsymbol{\theta}})_{\pi_0} = 10^{-3}(3.4591 - 1.8762) = (-0.0051 \div 0.0065) \times (-0.0020261),$$

and an analogous result holds for $\hat{\pi}_1$.

7 Discussion

This paper examined Fisher information loss in the ML estimators for a series of 2×2 tables with missing interior cell counts, given that the row margins are fixed and that the random column observations are the sum of two independent binomials. We obtained a simplified expression for the parameter covariance and showed it to be equal to the negative geometric mean variance loss of the parameters. We also presented an approximation to the expectation of the conditional variance of the missing cell counts and an expression for the ML parameter bias in 2×2 ecological tables.

Obviously, when using ML methodology for the analysis of incomplete tables, some simplifications are required to make the analysis tractable. In this paper, we assumed that the conditional probabilities are homogeneous across tables. An alternative that is straightforward to implement is to use a regression model where the probabilities depend on a set of covariates. In addition, the analysis was accomplished by assuming that the observed data likelihood function is the convolution of two binomial distributions. A valuable avenue of further inquiry would be to examine Fisher information loss in 2×2 ecological tables using beta-binomial distributions for the unobserved cell counts.

Acknowledgements

This research was supported by a grant from the Netherlands Organisation for Scientific Research (NWO), Division for Social Sciences (# 480-04-009).

References

Böckenholt, U. and W. R. Dillon (2000), Inferring latent brand dependencies, *Journal of Marketing Research* **37**, 72–87.

Cox, D. R. and E. J. Snell (1968), A general definition of residuals, *Journal of the Royal Statistical Society, Series* B **30**, 248–275.

Davison, A. C. and C. Semadeni (2004), Discussion of J. Wakefield, Ecological inference for 2 × 2 tables, *Journal of the Royal Statistical Society, Series* B **167**, 434–435.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977), Maximum likelihood from incomplete data via the *EM* algorithm, *Journal of the Royal Statistical Society, Series* B **39**, 1–38.

Golan, A., G. Judge and D. Miller (1996), *Maximum entropy econometrics. Robust estimation with limited data*, Wiley, New York.

Haber, M. (1989), Do the marginal totals of a contingency table contain information regarding the table proportions? *Communications in Statistics – Theory and Methods* **18**, 147–156.

Hamdan, M. A. and M. O. Nasro (1986), Maximum likelihood estimation of the parameters of the bivariate binomial distribution, *Communications in Statistics – Theory and Methods* **15**, 747–754.

IMAI, K., Y. LU and A. STRAUSS (2007), Bayesian and likelihood inference for 2×2 ecological tables: an incomplete-data approach, *Political Analysis*. doi:10.1093/pan/mpm017.

KING, G. (1997), *A solution to the ecological inference problem. Reconstructing individual behavior from aggregate data*, Cambridge University Press, Cambridge MA.

KING, G., O. ROSEN and M. TANNER (1999), Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research* **28**, 61–90.

KING, G., O. ROSEN and M. TANNER (2004), *Ecological inference. New methodological strategies*, Cambridge University Press, Cambridge MA.

KOCHERLAKOTA, S. and K. KOCHERLAKOTA (1992), *Bivariate discrete distributions*, Marcel Dekker, New York.

LOUIS, T. A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society, Series* B **44**, 226–233.

MAGNUSSEN, S. (2004), Prediction of 2×2 tables of change from repeat cluster sampling of marginal counts, *Canadian Journal of Forest Research* **34**, 1703–1713.

MCCULLAGH, P. (1984), On the elimination of nuisance parameters in the proportional odds model, *Journal of the Royal Statistical Society, Series* B **46**, 250–256.

MCCULLAGH, P. and J. A. NELDER (1992), *Generalized linear models*, 2nd edn, Chapman & Hall, London.

ORCHARD, T. and M. A. WOODBURY (1972), A missing information principle: theory and applications, *Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697–715.

PLACKETT, R. L. (1977), The marginal totals of a 2×2 table, *Biometrika* **64**, 37–42.

ROSEN, O., W. JIANG, G. KING and M. TANNER (2001), Bayesian and frequentist inference for ecological inference: the $r \times c$ case, *Statistica Neerlandica* **55**, 134–156.

SALWAY, R. E. and J. C. WAKEFIELD (2004), A common framework for ecological inference in epidemiology, political science and sociology, in: G. KING, O. ROSEN and M. TANNER (eds.), *Ecological inference. New methodological strategies*, pp. 303–332, Cambridge University Press, Cambridge MA.

STEEL, D. G., E. J. BEH and R. L. CHAMBERS (2004), The information in aggregate data, in: G. KING, O. ROSEN and M. TANNER (eds.), *Ecological inference. New methodological strategies*, pp. 51–68, Cambridge University Press, Cambridge MA.

TIAN, G., M. TAN and K. W. NG (2007), An exact non-iterative sampling procedure for discrete missing data problems, *Statistica Neerlandica* **61**, 232–242.

WAKEFIELD, J. (2004), Ecological inference for 2×2 tables (with discussion), *Journal of the Royal Statistical Society, Series* A **167**, 385–445.

WOODBURY, M. A. (1971), Discussion of H.O. Hartley and R. R. Hocking. The analysis of incomplete data, *Biometrics* **27**, 808–813.

Received: October 2007. Revised: October 2007.